

Geodesic acceleration and the small-curvature approximation for nonlinear least squares

Mark K. Transtrum

*Department of Bioinformatics and Computational Biology,
University of Texas M.D. Anderson Cancer Center, Houston Texas, U.S.A.**

James P. Sethna

*Laboratory of Atomic and Solid state Physics,
Cornell University, Ithaca, New York 14853, USA*

It has been shown numerically that the performance of the Levenberg-Marquardt algorithm can be improved by including a second order correction known as the geodesic acceleration. In this paper we give the method a more sound theoretical foundation by deriving the geodesic acceleration correction without using differential geometry and showing that the traditional convergence proofs can be adapted to incorporate geodesic acceleration. Unlike other methods which include second derivative information, the geodesic acceleration does not attempt to improve the Gauss-Newton approximate Hessian, but rather is an extension of the small-residual approximation to cubic order. In deriving geodesic acceleration, we note that the small-residual approximation is complemented by a small-curvature approximation. This latter approximation provides a much broader justification for the Gauss-Newton approximate Hessian and Levenberg-Marquardt algorithm. In particular, it is justifiable even if the best fit residuals are large, is dependent only on the model and not on the data being fit, and is applicable for the entire course of the algorithm and not just the region near the minimum.

*Electronic address: mkt26@cornell.edu

I. INTRODUCTION

In this paper we consider the problem of minimizing a scalar function whose form is the sum of squares

$$C(\theta) = \frac{1}{2} \sum_m r_m(\theta)^2, \quad (1)$$

where $r : \mathbf{R}^N \rightarrow \mathbf{R}^M$ is an M -dimensional vector function of N parameters, θ . We refer to $r(\theta)$ as the residuals and scalar function $C(\theta)$ as the cost. Functions of this form often arise in the context of data fitting and represent an important class of problems as evidenced by the large number of software packages dedicated to their optimization.

The structure of this particular problem lends itself to efficient optimization. In particular, consider the Hessian matrix of second derivatives, necessary to implement a quasi Newton method:

$$\begin{aligned} \frac{\partial^2 C}{\partial \theta_\mu \partial \theta_\nu} &= \sum_m \left(\frac{\partial r_m}{\partial \theta_\mu} \frac{\partial r_m}{\partial \theta_\nu} + r_m \frac{\partial^2 r_m}{\partial \theta_\mu \partial \theta_\nu} \right) \\ &\approx \sum_m \frac{\partial r_m}{\partial \theta_\mu} \frac{\partial r_m}{\partial \theta_\nu} \\ &= (J^T J)_{\mu\nu}, \end{aligned} \quad (2)$$

where in the last two lines we have applied the so-called Gauss-Newton or small-residual approximation and introduced the Jacobian matrix of first derivatives J . The approximation in the second line is usually justified by the hope that at a minimum of $C(\theta)$, the individual residuals are small, so that the Hessian is dominated by the contributions from the first term. Numerically, this approximation is advantageous since it allows one to implement a quasi-Newton method by calculating only the first derivatives of the residuals. This approximation comes at the cost of storing the derivative information for each residual individually, but this is rarely a bottleneck on modern computers. Consequently, the functional form of Eq. (1) effectively allows one to estimate the Hessian matrix with the same information used to calculate the gradient.

Applying a trust region method to the Gauss-Newton approximate Hessian results in the Levenberg-Marquardt algorithm[1–6] which iteratively updates the parameters according to

$$\delta\theta = - (J^T J + \lambda D^T D)^{-1} g, \quad (3)$$

where λ is an appropriately chosen Lagrange multiplier for the step bound $\delta\theta^T D^T D \delta\theta \leq \Delta^2$ and $g = J^T r$ is the gradient. Because Levenberg-Marquardt is a quasi-Newton method, it

usually has very good convergence properties. In particular, if the small-residual approximation is good, convergence can be super-linear to a local minimum. Furthermore, for well-constructed choices of λ and Δ , the method is globally convergent[3, 5].

Although the Levenberg-Marquardt algorithm has many desirable properties, it is not always ideal. Often data fitting problems have a cost function characterized by narrow, winding canyons. Although, asymptotically the algorithm may converge super-linearly, it may nevertheless spend an unreasonable amount of time navigating the winding canyon before it finally zooms into the minimum. This problem is typically more severe on problems with many parameters, which in turn are often more computationally expensive to evaluate and lack good parameter estimates to use as starting points. An improved optimization method which can find minima with fewer function evaluations (and especially fewer Jacobian evaluations) would be a welcome improvement.

In order to help improve the Levenberg-Marquardt algorithm, the authors previously proposed the inclusion of a geodesic acceleration term in the algorithm[7, 8]. This correction was derived using an information geometric interpretation of the least-square problem and was justified based on the empirically observed small-extrinsic curvatures of the relevant manifolds. In this paper, we will see that the geodesic acceleration correction can be understood as a generalization of the Gauss-Newton method extended to cubic order. Although other methods exist which utilize higher-order information, geodesic acceleration is complementary to these approaches as their primary motivation is to improve the estimate of the Hessian. By contrast, the geodesic acceleration assumes the Hessian estimate is adequate to proceed to higher-order.

In this paper, we derive the geodesic acceleration in a geometric independent way (section II) and prove that its inclusion in the Levenberg-Marquardt algorithm does not compromise its convergence properties (section III). From the explicit derivation in section II, we see that geodesic acceleration can be understood as a continuation of the small-residual approximation to higher-order terms. Indeed, the discarded terms are properly understood as the residuals coupled to the extrinsic curvatures of the *Model Graph* defined in references[7, 8]. The small-curvature approximation, therefore, provides additional justification for the Gauss-Newton approximation and the geodesic acceleration correction. As we argue in section IV, the small-curvature approximation is more useful than the small-residual approximation for many reasons. In particular, it is valid not only near a local minimum of the cost, but for

all parameter values; it is a property of the model and not the data being fit and so is valid even when the model cannot fit the data well; furthermore, numerical experiments on many models suggest the small-curvature approximation is nearly universally valid.

II. DERIVATION

In order to improve the efficiency of the Levenberg-Marquardt method, we propose modifying the step to include higher order corrections in a numerically efficient manner. To derive this correction, consider the minimization problem of finding the best residuals with a constrained step-size. We write the dependence of the residual on the shift $\delta\theta$ as

$$r(\theta + \delta\theta) = r + J\delta\theta + 1/2 \delta\theta^T K \delta\theta + \dots, \quad (4)$$

where J and K are the arrays of first and second derivatives respectively. We wish to minimize

$$\min_{\delta\theta} \left(r + J\delta\theta + 1/2 \delta\theta^T K \delta\theta \right)^2 \quad (5)$$

with the constraint that $\delta\theta^T D^T D \delta\theta \leq \Delta^2$. After introducing a Lagrange multiplier λ for the constraint in the step size, the minimization becomes

$$\min_{\delta\theta} \left(r + J\delta\theta + 1/2 \delta\theta^T K \delta\theta \right)^2 + \lambda \delta\theta^T D^T D \delta\theta. \quad (6)$$

By varying $\delta\theta$ we find the normal equations:

$$\begin{aligned} \sum_m J_{m\mu} r_m + \sum_{m\nu} (J_{m\mu} J_{m\nu} + r_m K_{m\mu\nu} + \lambda D_{m\mu} D_{m\nu}) \delta\theta_\nu \\ + \sum_{m\nu\alpha} (J_{m\nu} K_{m\mu\alpha} + 1/2 J_{m\mu} K_{m\nu\alpha}) \delta\theta_\nu \delta\theta_\alpha = 0, \end{aligned} \quad (7)$$

where we have explicitly included all the indices to avoid any ambiguity. Since we constrain the step size, it is natural to assume that $\delta\theta$ is small, and we seek a solution of Eq. (7) as a perturbation series around the linearized equation:

$$\delta\theta = \delta\theta_1 + \delta\theta_2 + \dots. \quad (8)$$

Let $\delta\theta_1$ be a solution of the linearized equation:

$$\begin{aligned} \delta\theta_1 &= -(J^T J + r^T K + \lambda D^T D)^{-1} J^T r \\ &\approx -(J^T J + \lambda D^T D)^{-1} J^T r, \end{aligned}$$

where in the second line we have made the usual Gauss-Newton approximation. We do not actually discard the term involving K , as it will help to cancel out a higher order correction later in the derivation. We therefore set

$$\delta\theta_1 = -(J^T J + \lambda D^T D)^{-1} J^T r, \quad (9)$$

which is the usual Levenberg-Marquardt step.

With this definition of $\delta\theta_1$, Eq. (7) becomes

$$\begin{aligned} & \sum_{m\nu} (J_{m\mu} J_{m\nu} + r_m K_{m\mu\nu} + \lambda D_{m\mu} D_{m\nu}) \delta\theta_{2\nu} \\ & + \frac{1}{2} \sum_{m\nu\alpha} J_{m\mu} K_{m\nu\alpha} \delta\theta_{1\nu} \delta\theta_{1\alpha} + \sum_{m\alpha} (r_m K_{m\mu\alpha} + \delta\theta_1^\nu J_{m\nu} K_{m\mu\alpha}) \delta\theta_{1\alpha} = 0. \end{aligned} \quad (10)$$

to second order, with the term $r_m K_{m\mu\alpha} \delta\theta_{1\alpha}$ the term neglected by the Gauss-Newton approximation at first order.

We now turn our attention to the second term in parentheses in Eq. (10). Using the definition of $\delta\theta_1 = -(J^T J + \lambda D^T D)^{-1} J^T r$, we can write

$$\begin{aligned} \sum_m r_m K_{m\mu\alpha} + \sum_{m\nu} \delta\theta_{1\nu} J_{m\nu} K_{m\mu\alpha} &= \sum_m r_m K_{m\mu\alpha} - \sum_{m\beta\nu} r_m J_{m\beta} (J^T J + \lambda D^T D)^{-1}_{\beta\nu} J_{n\nu} K_{n\mu\alpha} \\ &= \sum_{mn} r_m \left(\delta_{mn} - \sum_{\beta\nu} J_{m\beta} (J^T J + \lambda D^T D)^{-1}_{\beta\nu} J_{n\nu} \right) K_{n\mu\alpha}. \end{aligned}$$

Since this term is proportional to the residuals, r_m , it can be ignored using the usual small-residual arguments. However, we now make an appeal to geometric considerations by noting that $\delta_{mn} - \sum_{\beta\nu} J_{m\beta} (J^T J + \lambda D^T D)^{-1}_{\beta\nu} J_{n\nu} = P_{mn}^N$ is a matrix that projects vectors perpendicular to the tangent plane of the *Model Graph* as described in reference[8]. If the curvature of the model graph is small, then $P^N K \approx 0$ and this term can be neglected. We discuss the implications of this argument further in section IV.

Returning to Eq. (10), after ignoring the last term in parentheses, we find

$$\begin{aligned} \delta\theta_2 &= -\frac{1}{2} (J^T J + r^T K + \lambda D^T D)^{-1} J^T r'' \\ &\approx -1/2 (J^T J + \lambda D^T D)^{-1} J^T r'', \end{aligned} \quad (11)$$

where we have introduced the directional second derivative $r''_m = \sum_{\mu\nu} K_{m\mu\nu} \delta\theta_{1\mu} \delta\theta_{1\nu}$ and in the second line made the usual Gauss-Newton approximation to the Hessian, giving the formula first presented in [7]. This formula was originally interpreted as the second

order correction to geodesic flow on the model graph, and so we refer to it as the geodesic acceleration correction. By analogy, we refer to the first order term as the geodesic velocity. The full step is therefore given by:

$$\delta\theta = \delta\theta_1 + \delta\theta_2 \equiv v\delta t + 1/2 a\delta t^2. \quad (12)$$

As has been noted previously[7, 8], although the geodesic acceleration correction includes second derivative information at each step of the algorithm, its calculation is not computationally intensive. In particular, it only requires the evaluation of a directional second derivative, which is computationally comparable to a single evaluation of the residuals. Indeed, in the absence of an analytic expression, a finite difference estimate of the relevant second derivative can be found by a single function evaluation. In contrast, the Jacobian evaluation at each step is comparable to N function evaluations. Particularly for large problems, the computational cost of including the geodesic acceleration is negligible compared the other elements of the algorithm.

III. CONVERGENCE

In order to show that geodesic acceleration does not impair the convergence guarantees of the Levenberg-Marquardt algorithm, we must make a few additional modifications. We first note that an algorithm that selects Δ directly requires a solution of Eq. (5) given the step bound, i.e. find the value of λ corresponding to the step bound Δ . This so-called subproblem, can be solved accurately and efficiently for the case $K = 0$, using the methods described by Moré[9] and Nocedal and Wright[5]. When including geodesic acceleration however, such a simple solution does not exist. Indeed, the step size is no longer a monotonically decreasing function of λ . Furthermore, accounting for the contribution from the second term requires additional function evaluations, making an accurate solution computationally expensive.

Fortunately, convergence proofs for Levenberg-Marquardt do not require that this subproblem be solved accurately. We therefore content ourselves by approximately solving the problem as follows: We first require that $|\delta\theta_1| \leq \Delta$. This step can be satisfied easily using the algorithm in references[5, 9]. We next require that the relative contribution from the second order step be bounded

$$\frac{2|\delta\theta_2|}{|\delta\theta_1|} \leq \alpha \quad (13)$$

for some $\alpha > 0$ [17]. In practice we implement the requirement in Eq. (13) by rejecting all proposed steps for which it is not satisfied and decreasing Δ (or increasing λ) until an acceptable step is generated, recalculating $\delta\theta_1$ for the new Δ . This may appear inefficient since the method will on occasion reject steps that would have decreased the cost. However, this requirement adds stability to the algorithm, helping it to avoid undesirable fixed points with infinite parameter values as argued and numerically justified in references[7, 8, 10].

We need to make an additional assumption about the behavior of the model. Specifically, it is necessary to assume that the directional second derivative of the model is bounded: $|\sum_{\mu\nu} K_{m\mu\nu} u^\mu u^\nu| < \kappa$ for any parameter-space unit vector u and some positive constant κ . This assumption is necessary to guarantee that we can always find a step-size that satisfies Eq. (13). We do not anticipate this requirement to be a major restriction for the applicability of the algorithm as we discuss later in this section.

With this additional assumption, we can show that our modified Levenberg-Marquardt algorithm enjoys the same global convergence properties as original algorithm. Indeed, the proof is nearly identical to that of the unmodified Levenberg-Marquardt. Our proof here follows closely that of Theorem 4.5 in Nocedal and Wright[5]. First we define the model function $m_k(\delta\theta)$ by

$$\begin{aligned} m_k(\delta\theta) &= \frac{1}{2} (r_k + J_k \delta\theta)^2 \\ &= \frac{1}{2} |r_k|^2 + \delta\theta^T J_k^T r_k + \frac{1}{2} \delta\theta^T J_k^T J_k \delta\theta, \end{aligned} \tag{14}$$

and the reduction ratio ρ_k by

$$\rho_k = \frac{C(\theta_k) - C(\theta_k + \delta\theta_k)}{m_k(0) - m_k(\delta\theta_k)}. \tag{15}$$

We now define an algorithm for which we will prove convergence. This algorithm is analogous to algorithm 4.1 in reference[5] which we recover if we were to set $\delta\theta_2 = 0$ at each step.

Algorithm 1

Given $\hat{\Delta} > 0$, $\Delta_0 \in (0, \hat{\Delta})$, and $\alpha > 0$

for $k = 0, 1, 2, \dots$ **do**

 Calculate $\delta\theta_1$ and $\delta\theta_2$ as described above

 Set $\delta\theta_k = \delta\theta_1 + \delta\theta_2$

if $|\delta\theta_2| > \alpha|\delta\theta_1|/2$ **then**

$$\Delta_{k+1} = \frac{1}{4}\Delta_k$$

```

 $\theta_{k+1} = \theta_k$ 
else
  Evaluate  $\rho_k$  as in Eq. (15)
  if  $\rho_k < \frac{1}{4}$  then
     $\Delta_{k+1} = \frac{1}{4}\Delta_k$ 
  else
    if  $\rho_k > \frac{3}{4}$  and  $|\delta\theta_1| = \Delta_k$  then
       $\Delta_{k+1} = \min(2\Delta_k, \hat{\Delta})$ 
    else
       $\Delta_{k+1} = \Delta_k$ 
    end if
  end if
  if  $\rho_k > 0$  then
     $\theta_{k+1} = \theta_k + \delta\theta_k$ 
  else
     $\theta_{k+1} = \theta_k$ 
  end if
end if
end for

```

Before presenting our proof, first notice that by defining $\tilde{\theta} = D\theta$, the optimization problem in $\tilde{\theta}$ must at each step satisfy the bound $|\delta\tilde{\theta}_1| < \Delta$. Without loss of generality, we therefore assume that $D^T D$ is the identity. This assumption essentially replaces the Jacobian matrix with $\tilde{J} = JD^{-1}$. With this additional assumption, we now present a Lemma that will be useful in proving convergence.

Lemma 1

Suppose that at some point θ our function has (non-infinite) residuals r , Jacobian matrix J , and second derivative array K , which satisfy $|J^T J| < \beta$ and $|K_{m\mu\nu}u^\mu u^\nu| < \kappa$ for any parameter space unit vector u . Given positive constants $\alpha > 0$ and $\zeta > 1$, then if

$$\zeta\Delta \leq \frac{|g|}{\sqrt{\beta\kappa|r|/\alpha + \beta}}, \quad (16)$$

$$|\delta\theta_1| \leq \zeta\Delta, \quad (17)$$

then $|\delta\theta_2|/|\delta\theta_1| < \alpha/2$, where $g = J^T r$ is the function's gradient.

Proof

Let λ denote the Lagrange multiplier associated with the constraint in Eq. (17). Then $\zeta\Delta \geq |\delta\theta_1| = |(J^T J + \lambda)^{-1}g| \geq |g|/(\beta + \lambda)$, from which it follows that $\lambda \geq |g|/(\zeta\Delta) - \beta \geq \sqrt{\beta\kappa|r|/\alpha}$. Notice that since $|J^T J| \leq \beta$, the largest singular value of J must be less than $\sqrt{\beta}$. We therefore have

$$\begin{aligned} |\delta\theta_1| &= |(J^T J + \lambda)^{-1} J^T r| \\ &\leq |(J^T J + \lambda)^{-1}| |J^T| |r| \\ &\leq \frac{\sqrt{\beta}|r|}{\lambda}. \end{aligned} \tag{18}$$

Similarly,

$$|\delta\theta_2| \leq \frac{\sqrt{\beta}\kappa}{2\lambda} |\delta\theta_1|^2. \tag{19}$$

Combining these results gives us $|\delta\theta_2|/|\delta\theta_1| \leq \beta\kappa|r|/2\lambda^2 < \alpha/2$.

With this Lemma, we are prepared to prove our main result: that including geodesic acceleration does not affect the convergence properties of the Levenberg-Marquardt algorithm.

Theorem 1

Suppose that $|J^T J| \leq \beta$ for some positive constant β , that C is Lipschitz continuously differentiable in the neighborhood $S(R_0)$ for some $R_0 > 0$, and that $|\delta\theta_1| < \zeta\Delta_k$ for some constant $\zeta > 1$ at each iteration. Also assume that $m_k(0) - m_k(\delta\theta) \geq c_1|g_k| \min(\Delta_k, |g_k|/|J^T J|)$ for some positive constant $c_1 \in (0, 1]$ ($g_k = J_k^T r_k$), and that the second directional derivative of r_k in any unit direction u is bounded $|\sum_{\mu\nu} K_{m\mu\nu} u^\mu u^\nu| < \kappa$ for some positive κ . Then using Algorithm 1,

$$\liminf_{k \rightarrow \infty} |g_k| = 0. \tag{20}$$

Proof

The proof is nearly identical to that of Theorem 4.5 in reference [5]. We repeat the proof here in order to highlight the small-differences, leaving out the algebraic details.

With some algebraic manipulation, we obtain

$$|\rho_k - 1| = \left| \frac{m_k(\delta\theta_k) - C(\theta_k + \delta\theta_k)}{m_k(0) - m_k(\delta\theta_k)} \right|.$$

Using the same argument as in reference [5], we have

$$|m_k(\delta\theta_k) - C(\theta_k + \delta\theta_k)| \leq (\beta/2 + \beta_1) |\delta\theta_k|^2, \tag{21}$$

where β_1 is the Lipschitz constant for g on the set $S(R_0)$.

Suppose for contradiction that there is $\epsilon > 0$ and a positive index K such that

$$|g_k| \geq \epsilon, \quad \text{for all } k \geq K, \quad (22)$$

then we have

$$m_k(0) - m_k(\delta\theta_k) \geq c_1 |g_k| \min(\Delta_k, |g_k|/|J^T J|) \geq c_1 \epsilon \min(\Delta_k, \epsilon/\beta). \quad (23)$$

By the workings of Algorithm 1, we see that ρ is only calculated if $|\delta\theta_2| < \alpha|\delta\theta_1|/2$. In this case we have $|\delta\theta| = |\delta\theta_1 + \delta\theta_2| < |\delta\theta_1| + |\delta\theta_2| < (1 + \alpha/2)|\delta\theta_1| < (1 + \alpha/2)\zeta\Delta_k$. Each step therefore satisfies

$$|\delta\theta| < \gamma\Delta_k \quad (24)$$

where $\gamma = (1 + \alpha/2)\zeta > 1$.

Using Eqs. (24), (23), and (21), we have

$$|\rho_k - 1| \leq \frac{\gamma^2 \Delta_k^2 (\beta/2 + \beta_1)}{c_1 \epsilon \min(\Delta_k, \epsilon/\beta)}. \quad (25)$$

We now derive a bound on the right-hand-side that holds for all sufficiently small-values of Δ_k , that is, for all $\Delta_k \leq \bar{\Delta}$, where $\bar{\Delta}$ is defined as follows:

$$\bar{\Delta} = \min \left(\frac{1}{2} \frac{c_1 \epsilon}{\gamma^2 (\beta/2 + \beta_1)}, \frac{R_0}{\gamma}, \frac{\epsilon}{\zeta \sqrt{\beta \kappa} |r| / \alpha + \zeta \beta} \right). \quad (26)$$

As noted in reference [5], the R_0/γ term in this definition ensures that the bound in Eq. (21) is valid because $|\delta\theta_k| \leq \gamma\Delta_k \leq \gamma\bar{\Delta} \leq R_0$. The third term is necessary to show convergence when including geodesic acceleration.

Note that since $c_1 \leq 1$ and $\gamma \geq 1$, we have $\bar{\Delta} \leq \epsilon/\beta$. The latter condition implies that for all $\Delta_k \in [0, \bar{\Delta}]$, we have $\min(\Delta_k, \epsilon/\beta) = \Delta_k$, so from Eq. (25) and (26), we have

$$|\rho_k - 1| \leq \frac{1}{2},$$

following the logic in reference [5]. Therefore, if $\Delta_k \in [0, \bar{\Delta}]$, then $\rho_k > \frac{1}{4}$. Furthermore, the third condition in Eq. (26) implies that if $\Delta_k \in [0, \bar{\Delta}]$ then $|\delta\theta_2| < \alpha|\delta\theta_1|/2$ by Lemma 1.

Since, we have that if $\Delta_k \in [0, \bar{\Delta}]$, then $|\delta\theta_2| < \alpha|\delta\theta_1|/2$ and that $\rho_k > \frac{1}{4}$, by the workings of Algorithm 1, $\Delta_{k+1} \geq \Delta_k$ whenever Δ_k falls below the threshold $\bar{\Delta}$. Consequently, a

reduction in Δ_k (by a factor of $1/4$) can occur in the algorithm only if $\Delta_k \geq \bar{\Delta}$. We conclude that

$$\Delta_{k+1} \geq \min(\Delta_k, \bar{\Delta}/4) \quad \text{for all } k \geq K. \quad (27)$$

Suppose now that there is an infinite subsequence \mathcal{K} such that $\rho_k \geq 1/4$ for $k \in \mathcal{K}$. For $k \in \mathcal{K}$ and $k \geq K$, we have from Eq. (23) that

$$\begin{aligned} C(\theta_k) - C(\theta_{k+1}) &= f(\theta_k) - f(\theta_k + \delta\theta_k) \\ &\geq \frac{1}{4}[m_k(0) - m_k(\delta\theta_k)] \\ &\geq \frac{1}{4}c_1\epsilon \min(\Delta_k, \epsilon/\beta). \end{aligned}$$

Since C is bounded below, it follows from this inequality that

$$\lim_{k \in \mathcal{K}, k \rightarrow \infty} \Delta_k = 0,$$

contradicting Eq. (27). Hence no such infinite subsequence \mathcal{K} can exist, and we must have $\rho_k < \frac{1}{4}$ for all k sufficiently large. In this case, Δ_k will eventually be multiplied by $\frac{1}{4}$ at every iteration, and we have $\lim_{k \rightarrow \infty} \Delta_k = 0$, which again contradicts Eq. (27). Hence our original assertion, Eq. (22) must be false, proving the theorem.

Although Theorem 1 and its subsequent proof are nearly identical to Theorem 4.5 in Nocedal and Wright[5], we now briefly discuss its implication. In particular, the proof is that the algorithm converges to a point where cost has zero gradient. This does not automatically imply convergence in the parameters unless the Hessian matrix is bounded from below. In fact, the Hessian matrix can be very ill-conditioned, particularly for models with many parameters[11, 12]. Although the inferred parameters in such problems may be ill-conditioned, the convergence properties of the algorithm are nevertheless robust.

The requirement that $m_k(0) - m_k(\delta\theta) \geq c_1|g_k| \min(\Delta_k, |g_k|/|J^T J|)$ at each step of the algorithm in this context is given without motivation. However, just like the analogous theorem in reference[5], this requirement is necessary to guarantee that the parameter iterates do not accumulate at points with nonzero gradient. We note that in practice this condition is never explicitly checked.

The only new assumption of Theorem 1 beyond those for the standard algorithms[5] is the bound on the second derivative. Since we are utilizing second derivative information, this addition is not unexpected and guarantees that the new algorithm is well-behaved. It

is also clearly analogous to the bound on the first derivative used in the first algorithm ($|J^T J| \leq \beta$, equivalent to $|Ju| \leq \sqrt{\beta}$). The standard bound on the first derivative already excludes models where the cost diverges at interior points in the domain of the model as well as points like $C(\theta) \sim \sqrt{|\theta - \theta_0|}$ where the cost would stay finite although the derivative diverges; the new bound on the second derivative additionally excludes cases like $C(\theta) \sim |\theta - \theta_0|^{3/2}$, which should not arise often in practice. In our experience, many models do have costs which become singular at unphysical points (*i.e.*, positive-only parameters set less than zero), which can be addressed by an appropriate reparameterization of the model (*i.e.*, shifting to log parameters). When this can be done, it is also likely to improve the convergence rate of the algorithm.

We have here shown convergence of geodesic acceleration for an algorithm that belongs to the broad class of trust region methods that operate by specifying a step bound Δ . Originally Levenberg-Marquardt was proposed as an algorithm that heuristically selected λ directly rather than implicitly through Δ [1, 2]. For an algorithm that directly selects λ , one can similarly show convergence[3]. The proof that geodesic acceleration does not impair the convergence of this class of algorithm is almost identical to that of the original theorem, just as the proof of Theorem 1 above is almost identical to that of Theorem 4.5 in reference[5]. We do not give the rigorous proof here, but merely note that under the same additional assumptions, convergence can be shown for these methods as well.

IV. SMALL-CURVATURE APPROXIMATION AND RELATION TO OTHER METHODS

In deriving the geodesic acceleration in section II, we observed that the neglected terms were each proportional to the residuals and could be neglected in the small-residual approximation. We also noted that the neglected terms were also proportional to the extrinsic curvature on the model graph as described in reference[7, 8]. Indeed, the geodesic acceleration was originally understood as a small-curvature approximation rather than a small-residual approximation. These two approximations are complementary; in general only one of the two needs to be valid in order for the approximations to hold.

In deriving the geodesic acceleration, the connection between the small-residual and small-curvature approximation only became apparent when considering cubic order terms. How-

ever, the equivalence of the two approximations can be seen in the neglected terms of the Gauss-Newton Hessian without considering the geodesic acceleration. At a fixed point of the cost, the residuals are perpendicular to the model manifold (a fact used elsewhere to justify a scale-free measure of convergence[13]), i.e. $r_n \approx \sum_m r_m P_{mn}^N$. Thus, as the algorithm approaches a minimum, the neglected term can be written as

$$\sum_m r_m \frac{\partial^2 r_m}{\partial \theta_\mu \partial \theta_\nu} \approx \sum_{mn} r_m P_{mn}^N \frac{\partial^2 r_n}{\partial \theta_\mu \partial \theta_\nu}. \quad (28)$$

As Eq. (28) makes clear, near the best fit the nonlinear contribution to the Hessian includes only components perpendicular to the tangent plane and is thus proportional to the extrinsic curvature.

The implication of Eq. (28) is that the Levenberg-Marquardt algorithm may attain very good convergence rates even with large residuals so long as the extrinsic curvature is sufficiently small. However, the approximation in Eq. (28) is only valid near a fixed point of the cost, just as is the small-residual approximation is only valid near the minimum. The advantage of identifying the equivalence of the small-residual and small-curvature approximations is that the latter is likely to have much wider applicability. In particular for data fitting, the small-curvature approximation is a feature of the model rather than the data. Thus, although the validity of the small-residual approximation cannot be identified without finding the best fit, the small-curvature approximation can be. Furthermore, the small-curvature approximation is likely to be an excellent approximation for most models. In particular, several examples in references[8, 14] exhibit extrinsic curvatures many orders of magnitude smaller than the magnitude of the bare nonlinearities.

Although Eq. (28) is only valid near a minimum, the approximation that led to the geodesic acceleration is valid over a much larger parameter range, and it is for this reason it is likely to be a useful modification. The problem with Levenberg-Marquardt is not that its asymptotic convergence rate is poor (super-linear convergence is satisfactory in most cases), but that it may spend an unreasonable amount of time navigating a narrow canyon before that convergence rate is realized. As originally suggested, the geodesic acceleration can speed up this process since it describes the curvature of the canyon. The algorithm can find the minimum more quickly by following the canyon with a sequence of parabolic steps, rather than linear steps. A numerical demonstration of this will not be given here, as it has already been shown elsewhere[7, 8, 10].

There are many other methods which include second derivative information in order to improve the Levenberg-Marquardt algorithm. These approaches all try to improve the quality of the Hessian estimate in some way[5, 15, 16]. Thus, although geodesic acceleration may appear superficially similar to these approaches, it is actually quite different. The philosophy behind geodesic acceleration is to extend the small-residual/curvature approximation to higher order terms rather than estimate the neglected terms. Of course, the utility of such of an approach is ultimately measured by performance on real world problems. As has been shown elsewhere, geodesic acceleration can be very helpful on large problems for which Levenberg-Marquardt spends unreasonable time searching for a minimum before zooming into the best fit.

V. CONCLUSION

In this paper we have studied the geodesic acceleration correction to the Levenberg-Marquardt algorithm, originally suggested in references[7, 8]. We have derived the correction without the use of differential geometry and shown that it can be interpreted as an extension of the small-residual approximation used to estimate the Hessian matrix of a sum of squares. We have also seen that the small-residual approximation is complemented by the small-curvature approximation, which is likely to be applicable under much more general circumstances.

We have seen that with just a few minor modifications, the geodesic acceleration correction can be incorporated into a standard Levenberg-Marquardt routine with minimal computational cost and without sacrificing its convergence properties. Numerical experiments given elsewhere[7, 8, 10], suggest that the benefit/cost ratio of this improvement can be substantial on many optimization problems.

The authors would like to thank Cyrus Umrigar, Peter Nightingale, Saul Teukolsky, and Ben Machta for helpful conversation. This work is partially supported by the TCGA Genome Data Analysis Center (GDAC) grant (MKT) and the Cancer Center Support Grant at the University of Texas MD Anderson Cancer Center (U24 CA143883 02 S1 and P30 CA016672)

(MKT) and by NSF grant number DMR-1005479 (JPS).

- [1] K. Levenberg: *Quart. Appl. Math* **2** (1944) 164
- [2] D. Marquardt: *Journal of the Society for Industrial and Applied Mathematics* **11** (1963) 431
- [3] M. Osborne: *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics* **19** (1976) 343
- [4] J. Moré: *Lecture notes in mathematics* **630** (1977) 105
- [5] J. Nocedal, S. Wright: *Numerical optimization*: Springer (2000)
- [6] W. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery: *Numerical recipes: the art of scientific computing*,: Cambridge University Press (2007)
- [7] M. K. Transtrum, B. B. Machta, J. P. Sethna: *Physical Review Letters* **104** (2010) 1060201
- [8] M. K. Transtrum, B. B. Machta, J. P. Sethna: *Physical Review E* **83** (2011) 036701
- [9] J. Moré, D. Sorensen: *SIAM Journal on Scientific and Statistical Computing* **4** (1983) 553
- [10] M. Transtrum, J. Sethna: *Arxiv preprint arXiv:1201.5885* (2012)
- [11] J. Waterfall, F. Casey, R. Gutenkunst, K. Brown, C. Myers, P. Brouwer, V. Elser, J. Sethna: *Physical Review Letters* **97** (2006) 150601
- [12] R. Gutenkunst, J. Waterfall, F. Casey, K. Brown, C. Myers, J. Sethna: *PLoS Comput Biol* **3** (2007) e189
- [13] D. Bates, D. Watts: *Technometrics* **23** (1981) 179
- [14] D. Bates, D. Watts: *J. Roy. Stat. Soc* **42** (1980) 1
- [15] J. Dennis Jr, D. Gay, R. Walsh: *ACM Transactions on Mathematical Software (TOMS)* **7** (1981) 348
- [16] R. Fletcher: *Practical methods of optimization, Volume 1*: Wiley (1987)
- [17] In practice $\alpha \approx 0.75$ is a robust and efficient choice